Moderation and Consistency of Teacher Judgement: Teachers' Views

Stephen Connolly,
Val Klenowski and
Claire Wyatt-Smith

Abstract
Major curriculum and assessment reforms in Australia have generated research interest in issues related to standards, teacher judgement and moderation. This article is based on one related inquiry of a large-scale Australian Research Council Linkage project conducted in Queensland. This qualitative study analysed interview data to identify teachers' views on standards and moderation as a means to achieving consistency of teacher judgement. A complementary aspect of the research involved a blind review that was conducted to determine the degree of teacher consistency without the experience of moderation. Empirical evidence was gained that most teachers, of the total interviewed articulated a positive attitude towards the use of standards in moderation and perceived that this process produces consistency in teachers' judgements. Context was identified as an important influential factor in teachers' judgements and it was concluded that teachers' assessment beliefs, attitudes and practices impact on their perceptions of the value of moderation practice and the extent to which consistency can be achieved.

Introduction
The recent move towards high stakes national student testing and national reporting of school outcomes to meet public accountability demands in Australia has intensified interest in the extent to which teachers' judgements are reliable and consistent. Globally governments have responded to the international comparative analyses of student achievement data, as reflected in the Program for International Student Assessment (PISA) or Trends in Mathematics and Science Study (TIMSS), by implementing standards-driven curriculum and assessment reform. Australia is no exception with plans for a National Curriculum and ongoing national testing or the National Assessment Program - Literacy and Numeracy (NAPLAN). These changes have challenged State efforts to maintain the emphasis on assessment to promote learning while fulfilling accountability demands.

The study reported here focused on the introduction of social moderation practice, in the context of standards-referenced assessment, to middle school teachers (Years 4–9) who had no prior experience of this practice. This inquiry is part of a large-scale Australian Research Council (ARC) Linkage project, *Investigating Standards-Driven Reform in Assessment in the Middle Years of Schooling*, which was conducted in Queensland[i]. The researchers explored teachers' views of their experience of the social moderation process and their perceptions regarding the consistency and comparability of the judgements of student achievement as an outcome of these moderation processes.

More detail relevant to the policy context of this study is now given to set the scene for the research. The critical review of the literature on teacher judgement within large-scale systems of assessment has revealed a paucity of research conducted specifically on teachers' perceptions regarding the consistency of teacher judgement. The research aims, research design and methodology adopted are outlined. The analysis and findings are then discussed with the implications for policy and practice in this time of national curriculum and assessment change made explicit.

Context

The Australian Curriculum, Assessment and Reporting Authority (ACARA) is responsible for the development of a national curriculum and national achievement standards. Constitutionally, however, the power to decide on school curriculum still resides with state governments rather than with the federal government. This division of powers has impeded past developments towards a national curriculum, and publication of student results, with this recent endeavour being no exception to this trend.

Accountability testing in Australian public education policy has gained prominence since 2008 when Australia's National Assessment Programme – Literacy and Numeracy (NAPLAN) was extended from Years 3, 5 and 7 to include students in Year 9. At the same time Commonwealth funding legislation moved from being state-based to a national testing system. At the present time students in Years 3, 5, 7 and 9 sit national tests in reading, writing, spelling, grammar and punctuation and numeracy. However, despite these developments in national testing there has been no direct link to a national curriculum.

The Ministerial Council for Education, Employment and Youth Affairs (MCEETYA) in April 2009 announced a decision to develop a system for comparing the performance of schools using NAPLAN results and other sources of data. This step towards greater transparency has resulted in the *My School* website (www.myschool.edu.au). Such measures have impacted on state governments that are now keen to raise standards as represented by the results of NAPLAN tests.

These major educational reforms have led each state of Australia to develop local systems to support teachers. In Queensland teacher assessment has historically been seen as a source of dependable results through moderation practice. The Queensland Studies Authority (QSA) has recognised and supported the development of teacher assessment and moderation practice. These efforts are now contested as political pressures related to the national testing, and national partnership funding arrangements tied to the performance of students at or below minimum standards are being implemented. For 40 years high-stakes assessment of senior secondary students (Years 11 and 12, with students aged 17-18 years) has involved school-based assessments externally moderated using defined standards. At the heart of the Queensland senior assessment model is social or consensus moderation (as distinct from statistical). External standards-referenced moderation has been routinely undertaken in these year levels as a main means to ensure accountability and to maintain public confidence in teacher judgement. Teacher moderated assessment is then "moderated' again statistically using a core skills test for the purpose of determining a student's university entrance score, referred to as an Overall Position (OP). It is emphasised that the teacher "moderated" standards-referenced judgements are not changed as a consequence of core skills test results.

In the primary and middle schools (Prep to Year 10) teachers have not until recently been required to use standards for assessing and grading purposes, nor have they had to undertake inter or intra school moderation as part of system efforts to support consistency of teacher judgement. The first 11 years of schooling (P-10) until recently were considered low-stakes by the Queensland Studies Authority, teachers and policy officers. Up until 2008 during these

years of schooling there has been an absence of formal checks and balances in place at system level to confirm the validity and reliability of teacher judgement. Recently the Queensland Curriculum, Assessment and Reporting Framework (QCAR) of Essential Learnings, A-E standards and a common reporting framework, to promote consistency of teacher judgement, were introduced by QSA. The Queensland Comparable Assessment Tasks (QCATs) were also introduced in Years 4, 6 and 9 for English, Mathematics and Science. A generic set of descriptors (A – E standards) of the expected quality of student work was also developed for each Key Learning Area (KLA). These standards for Essential Learnings (ELs), provide a common language for teachers to use in assessing student work (QSA, 2007). The standards used to assess the QCATs are specific to the assessment task and align with the generic KLA standards. The QCAT standards are also expressed as alphabetic A-E descriptors.

The QSA in introducing the QCATs and task related standards aimed to promote teachers' professional learning, develop their assessment capacity and provide them with an opportunity to participate in system level standards-referenced reporting. It was expected that by using the standards and engaging in moderation teachers would present more meaningful reports, arrive at more consistent judgements and engage with assessment as a learning process. To achieve consistency in the application and use of the standards requires that these are explicitly stated and exemplified

Teachers in this study were provided with the *Guide to Making Judgements* developed by QSA to assist them in assessing and grading the QCATs. The QSA in addition to providing the *Guide to Making Judgements* provided annotated examples for each A-E grade for each QCAT to assist teachers in using these standards to make a judgement. The teachers consulted the materials and documents provided by QSA to identify the qualities in the student work that would help them determine the grade to be awarded. Many P-10 teachers did not have experience of moderation and the senior secondary approach had not had a backwash effect to primary and middle years of schooling. At the time of this study QSA was designing, trialing and developing the QCATs, associated moderation practices and introducing the ELs. It needs to be emphasised that these teachers were using QCATs and defined standards for the first time and participating in moderation processes in the context of a trial.Theyhad not received training in either the use of the tasks and accompanying defined standards and many had no experience with moderation practices. At the same time they were also implementing the ELs.

Teacher Judgement

A critical review of the research pertaining to teacher judgement reveals that teachers draw on multiple sources of knowledge and evidence when making judgements (Cooksey, Freebody & Wyatt-Smith, 2007; Davison, 2004) and that the use of standards and criteria alone will not result in consistency of teacher judgements (Wyatt-Smith, Klenowski and Gunn, 2009). Sadler (2009) has identified context as an important factor in teacher use and interpretation of criteria and standards. Teachers and students will have different interpretations of standards and it is possible for a teacher to have a different interpretation of the same standards in different contexts (Sadler, 2009). Given the unique historical and cultural circumstances of teachers' and students' assessment experiences these studies have concluded that a variety of influences and knowledges impact on teacher judgement.

In the three-year large-scale Australian study of teacher judgement in middle schooling, Cooksey, Freebody, and Wyatt-Smith (2007) reported high levels of variability in teachers' notions of quality and also unearthed the range of factors that shape how judgements are

reached (Wyatt-Smith, 1999; Wyatt-Smith & Castleton, 2004). Cooksey, Freebody and Wyatt-Smith (2007) developed the notion of assessment as *judgement-in-context* in their study of the ways in which teachers use assessment evidence, and their knowledge of their students, in a given assessment framework. These authors compared analyses of teachers' assessments using the "individual judgment processes that they normally apply in their classrooms versus those made using an externally provided set of national standards ('benchmarks')" (p. 402). The study revealed "… some significant conclusions about the highly variable influence that context [local and system] exerts on teachers' judgmental consistency and agreement" (p. 429). A variety of teachers' judgement systems was identified. The 20 teacher participants were representative of the full range of cultural, linguistic, socioeconomic and academic factors. The level of diversity was a particularly salient finding that revealed different teacher judgement systems across various sites "using either their own native assessment systems or a national system of benchmarks mandated by the state" (Cooksey et al., 2007, p. 404).

Further research has identified a duality in the process of teacher judgement attributable to the influence, to various degrees, of "criterion-referenced or construct-referenced" (Davison (2004, p. 308) factors embedded in teacher assessment beliefs, attitudes and practices. Davison (2004) developed a framework that incorporates the dual perspectives on a cline with criterion-referenced and construct-referenced dispositions at either end of the scale which "provides a mechanism to describe more systematically the effects on teachers of different sorts of assessment approaches, including norm, construct and criterion-referenced, and the interaction of these frameworks with their professional knowledge" (p. 324). Davison's (2004) findings correspond with those of Cooksey et al. (2007) in that personal approaches to assessment are juxtaposed with prescribed assessment systems and both impact on the degree of consistency in teachers' judgements.

Social moderation involves the use of teacher judgement with standards referencing or "the direct apprehension and comparison of standards" (Maxwell, 2009: 459). This judgement practice involves personal comparison and alignment of assessor judgements. It is a participative process that respects the professionalism of teachers as assessors (Maxwell, 2009) who meet to consult one another in considering the judgements and to achieve consensus on the standards or grades awarded. Consistency of teacher judgement is achieved when teachers agree on the standard or grade awarded to the student's response to a task. If consistency in the application of these standards exists, then it can be said that there is comparability across the assessment grades awarded. Moderation involves processes of consultation, negotiation and application of standards to achieve consensus or agreement (Klenowski & Adie, 2009). Teachers usually meet in schools or other locations to discuss the quality of the student work with reference to the standards. It is also possible, as was trialled in the larger context of this study, for teachers to synchronously meet and moderate in an online situation using IT software (Adie, 2010). In this study the WebEx© Meeting Centre, a commercial web-conferencing software package was used. In these online meetings an individual teacher's grades were moderated by a group of teachers to achieve consensus.

Design of the study

Aims and Rationale

The aims of this research were to study teachers' views about the comparability and consistency of their judgements as a consequence of participating in social moderation

practices, many of whom were doing so for the first time. The researchers were interested to explore teachers' perspectives of their use of standards as practised in the moderation processes of QCATs and their viewpoints on whether this resulted in greater consistency, comparability and alignment of assessor judgements.

The larger ARC Linkage project collected a rich qualitative data set from interviews of school principals, Heads of Departments (HODs) and teachers. The interviews were approximately 20 minutes in duration and were designed to generate responses to questions about the process of moderation and the teacher's use of standards. The research focus question for the study on which this article is based was: *What are teachers' perspectives of the use of standards and moderation as a means to achieving consistency of teacher judgement?*

Further research analysis questions stemmed from this main question. These were:

- *When teachers speak about moderation do they relate a positive, neutral or negative attitude towards the concept of the use of standards in moderation practices?*
- *How do teachers view the use of standards in moderation? Are they supportive, against, or undecided?*

Qualitative Research Approach

A qualitative approach to research was adopted in order to understand teachers' views of the relationship of the use of standards and moderation to consistency of teacher judgement. The study employed the results of a blind review process and semi-structured interviews. The semi-structured interview questions were derived from research questions and topics emerging from previous related studies (see for example, Klenowski, 2006, 2007; Sadler, 2005; Wyatt-Smith & Castleton, 2005). Data was drawn from participant interviews that took place prior to engagement with moderation processes, (pre-moderation), and again after engagement with moderation processes (post-moderation). This comparison of perspectives from these stages of moderation was considered necessary to identify teachers' views and understanding of the relationship of moderation practice to consistency of teacher judgements.

Data Collection

Data collection for this study involved semi-structured interviews (N 113) that were conducted, pre and post moderation practice. There were 67 participants in total, teachers or Head of Departments (HODs) (who also teach) and two principals. Interviewees were from 24 different schools, 15 state, 6 Catholic, 1independent and 2 special schools. These schools were located in remote, rural, and metropolitan areas from Far North Queensland to Brisbane, and the Gold Coast, and inland to Central Queensland.

Prior to the moderation process the participants selected a representative sample of assessed QCATs from their year level (years 4, 6 or 9) and subject domain (English, Maths or Science). The participants had assessed these tasks using the *Guide to Making Judgements* and the QCATs *Sample Responses* supplied by QSA. At this stage the participants were interviewed and asked about their assessment processes and their experiences of assessing the QCATs using the guidelines and support materials provided.

The teachers met with other teachers, either in their own school or another location, or participated in a synchronous online ICT meeting using the WebEx© Meeting Centre. During

the face-to-face moderation meetings the awarded grades of the assessed QCATs were compared, analysed and considered. During the online moderation meetings, a similar moderation process was followed while the teachers used the WebEx© software to highlight the reasons why they had awarded the grades to particular samples of work. Following both modes of moderation meetings participants were again interviewed and questioned about the process they had experienced and their use of the standards.

These two stages of interview data collection (pre- and post-moderation) were repeated the following year. Although the participant cohort had mostly changed, a substantial amount of second round data was collected. This extensive data collection process yielded 113 individual interviews inclusive of interviewees who had been interviewed more than once as a consequence of either their participation in both modes of moderation (face-to-face and online) or their pre- and post-moderation interviews. This vast data set of transcripts provided the primary data set for this study.

Blind Review

In an effort to determine the degree of teacher consistency in judgement, without the experience of moderation, a blind review of marked responses was conducted and the results analysed. The schools involved were purposively selected and represented the state, independent and Catholic sectors, remote, regional and metropolitan regions, socio economic, cultural and linguistic diversity. During this exercise a total number of 316 student responses to the English, Maths and Science QCATs were assessed by teachers in 9 different schools, using the standards A–E. The grades awarded were recorded. The graded responses were then cleaned of any marks or comments and sent to other schools for the blind review process. Teachers in these other schools graded the work again.

The grades awarded by both sets of teachers were recorded, processed and compared in spreadsheets. When the grades were compared there were a total number of 103 totally consistent grades from 316 marked and reviewed student responses. Of these 316 original responses 24 were identified as containing anomalies that rendered them partially, or in some cases wholly, unsuitable for inclusion in the comparable categories of either totally consistent or varying by at least one standard. The anomalies occurred when teachers did not make a judgement that awarded a distinct grade. For example some teachers graded work as A/B or C/D.

The 103 grades that were consistently graded by both sets of teachers comprised: 34 English, 38 Maths and 31 Science QCATs. The number of assessed responses that differed by one grade consisted of: 33 English, 56 Maths and 60 Science .The number of blind reviewed QCATs that differed by two grades involved: 7 English, 17 Maths and 13 Science. Those that were least comparable and differed by three grades were 3 Maths and 2 Science, there were none recorded in English for this category. When both sets of results are compared for the blind review only 35% of grades were totally consistent. The next largest category comprised responses of one grade difference. Table 1 illustrates all the categories and the spread of results:

Table 1

Blind review results of comparison of teacher judgement.

Consistency	English	Maths	Science	Totals
Totally Consistent	34	38	31	103
1 Grade Difference	33	56	60	149
2 Grades Difference	7	17	13	37
3 Grades Difference	0	3	2	5
Totals	74	114	106	294

These results of the blind review are illustrated in graphic form in Figure 1:

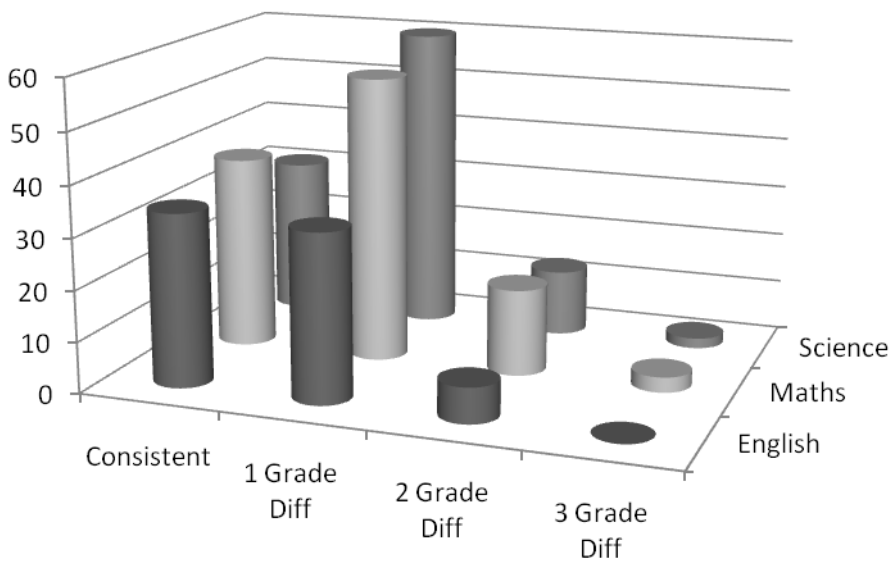

Figure 1. Graph of teacher consistency from blind review analysis

The blind review exercise demonstrated that teachers tended to arrive at comparable judgements. The most comparable was designated as a totally consistent result (N103), the next comparable category was one grade difference (N 149). That is, a ratio of 2:3, or expressed as percentages of the total number of included responses (294), 35% and 51% respectively, with two grade differences being 12% and three grade differences 2%. English teachers appear most consistent in their judgements with the results from the KLA of Science appearing to be categorised as least consistent. Maths recorded the highest numbers of two and three grade differences, which although somewhat paradoxical, reflects the nature of the 'authentic' task and the requirement for students to reflect on and communicate how they solved the inherent mathematical problems. It is worth emphasising here that the QCATs were assessing constructs that were included in the newly developed Essential Learnings (ELs). Constructs such as thinking and reasoning and discussion of the choice of strategies to complete particular mathematics tasks, for instance, were unfamiliar to teachers and students. Due to the pressures of timelines, budgets and product expectations the alignment of the ELs with the constructs of the QCATs was not always possible because the ELs were being introduced at the same time as the tasks were being trialed, developed and administered.

Data Analysis, Categories and Codes

A qualitative research paradigm was adopted. The interview transcripts were analysed using qualitative data analysis techniques of organising, matching, coding, identifying patterns and themes. The NVivo software package was used to sort the data into two main interview data sets of pre-moderation and post-moderation. Within these two sets two sub-categories were created for those interviewees involved in ICT moderation: ICT pre-moderation and ICT post-moderation.

Qualitative data analysis (Creswell, 2008) involved the researchers reading the transcripts and coding them individually. First they read the interview transcripts to ascertain the general content of the interviewees' responses and to familiarise themselves with overall content and tone of the interviews. Transcripts were then read again with the research questions used as a lens to highlight responses that related to the specific focus of this sub-study. This time the texts were scanned for participants' articulated responses that related to the research questions. Sections of transcripts were highlighted if they were evaluated as providing evidence of teachers' views about the relationship of moderation and consistency of teacher judgement. A pre-determined set of descriptors (Bazeley, 2007; Lincoln & Guba, 1985; Miles & Huberman, 1994; Saldaña , 2009) was identified for use in this coding exercise. This was "a formal inductive process of breaking down data into segments or data sets which can then be categorised, ordered and examined for connections, patterns and propositions that seek to explain the data" (Simons, 2009, p. 117). Once the data sets were coded they were then checked for accuracy and reliability. After the coding of the transcripts was checked, compared and analysed, themes aligned to the research questions were identified.

The set of descriptors developed from the research analysis questions provided the means for categorising participants' perspectives further. The resultant three categories of positive, neutral and negative attitudes were used to analyse the pre-moderation and post-moderation sets and sub-sets forming a total of 12 descriptors. The three categories are now described in detail.

Positive Attitude
Transcript excerpts were coded into this category of positive attitude if interviewees were favourable towards the idea that the use of standards in moderation would provide for consistency of teacher judgements. These excerpts indicated a general belief in the moderation process or a more specific attitude that consistency of teacher judgement was an outcome. Excerpts indicated a genuine belief that the use of standards in moderation is an appropriate process for achieving consistency in teachers' judgements and it is a fair and equitable process for agreeing on the award of student grades.

Neutral Attitude
Interview excerpts were coded as neutral if there were no distinctive qualities revealing any attitude in general towards: moderation, standards or the use of standards in moderation, consistency of teacher judgement, or any combination of these. Responses that indicated little or no knowledge of these concepts or claimed neutrality on the grounds of being insufficiently informed were also placed into this category. Others included some positive attitudes as well as some negative attitudes, or claims of being undecided. Overall the neutral attitude comprised those who were not sure, not informed, not concerned, uninterested or undecided.

Negative Attitude

Interviewees' responses that indicated a sceptical regard for moderation or standards and/or the idea of consistency of teacher judgement as an attainable goal either with or without the moderation process were categorised as negative. Disapproving views about using standards in moderation, either entirely or partially, and a pessimistic view of the likelihood that consistency in teachers' judgements would result, were categorised as negative.

The three categories of positive, neutral and negative attitudes were used to analyse the pre- and post-moderation data sets, which resulted in 12 pre-determined descriptors (Table 2).

Table 2

Coding categories

	Descriptor category	**Group**	**Code**
1	Positive attitude	pre-moderation	pospre
2	Positive attitude	post-moderation	pospost
3	Neutral attitude	pre-moderation	neupre
4	Neutral attitude	post-moderation	neupost
5	Negative attitude	pre-moderation	negpre
6	Negative attitude	post-moderation	negpost
7	ICT – Positive attitude	pre-moderation	ictpospre
8	ICT – Positive attitude	post-moderation	ictpospost
9	ICT – Neutral attitude	pre-moderation	ictneupre
10	ICT – Neutral attitude	post-moderation	ictneupost
11	ICT – Negative attitude	pre-moderation	ictnegpre
12	ICT – Negative attitude	post-moderation	ictnegpost

These descriptors were used to code the teachers' views. The tables that follow present the coded data with the number of comments in each category broadly suggesting the trends in the interviewees' views. The codes used are included so that 27 x pospre is the code used to express that 27 interviewees indicated a positive attitude in the pre-moderation interview. This analysis is then supplemented with rich qualitative data to explain the most important and frequently mentioned perspectives. The discussion section presents the themes related to the teachers' views enhanced with excerpts from the transcripts.

Table 3

Teachers' Views Coded

	Positive attitude	**Neutral attitude**	**Negative attitude**
Pre-moderation	27 x pospre	24 x neupre	5 x negpre
Post-moderation	35 x pospost	8 x neupost	4 x negpost
ICT Pre-moderation	3 x ictpospre	4 x ictneupre	0 x ictnegpre
ICT Post-moderation	10 x ictpospost	0 x neupost	0 x ictnegpost

Several factors need to be made explicit in how these results were formulated. For example:

- 17 were interviewed only pre-moderation,
- 2 were interviewed only post-moderation,
- 1 was interviewed only pre-ICT moderation,

- 2 were interviewed only post-ICT moderation,
- 2 were interviewed pre and post-moderation but no data was coded using the descriptors.
- 1 was interviewed post-moderation interview but no data was coded using the descriptors.
- 7 participants were interviewed pre and post-moderation as well as pre-ICT and post-ICT moderation.

Despite these factors Table 3 indicates more interviewees held a positive attitude towards the concept of using standards in moderation to provide for teacher consistency of judgement than the other two categories. The listed factors do however limit the inferences that can be drawn from the results.

Table 4 compares the coding of interviewees' pre-moderation attitude descriptor with the post-moderation attitude descriptor. That is, where interviews were conducted with the same person pre-moderation and post-moderation (N 33 + 7 ICT = 40) the descriptor code of the first interview is compared with that of the second interview. Any changes in attitude and possible trends in the teachers' responses are identified.

Table 4

Changed attitudes from pre to post-moderation interview

	CHANGED		**NO CHANGE**	
Number of teachers	**changed from pre-moderation attitude**	**to post-moderation attitude**	**articulated a pre-moderation attitude of**	**maintained attitude in post-moderation**
2	Positive	negative		
4	Positive	neutral		
16			positive	positive
12	Neutral	positive		
2	Neutral	negative		
2			neutral	neutral
2	Negative	positive		
0	Negative	neutral		
0			negative	negative

After the pre-moderation interview with this group there were 22 interviewees (55%) coded under the positive attitude descriptor, 16 (40%) coded as neutral and 2 (5%) coded as holding a negative attitude. Of the 55% positive, 5% changed to negative and 10% changed to neutral; 40% showed no change in attitude and remained positive. Of the 40% coded as neutral 30% changed to positive, 5% changed to negative and 5% maintained a neutral attitude. In the negative category 5%, the total number for this category, changed to positive, thus there were no other changes for this category. This shows a trend towards the positive attitude descriptor.

From this analysis 20% of teachers (8) expressed a reduction in confidence regarding the relationship between the use of standards and moderation and consistency of teacher judgement by the post-moderation interview.

From the 40 interviewees mentioned above only 7 were participants in the ICT moderation process. When those 7 were separated out from the above results 1 had changed from positive to neutral, 2 had no change from positive and 4 had changed from neutral to positive.

This analysis reflects similar trends for the non-ICT and the ICT-moderated responses although the number of participants involved is very small. Teachers involved with the ICT trial were identified as having a positive attitude however it is worth noting that one teacher did show an opposite trend that correlates with the earlier analysis (Table 4).

Figure 2 represents an alternative view of the data analyses of Table 4. This chart illustrates the trend towards the positive attitude descriptor.

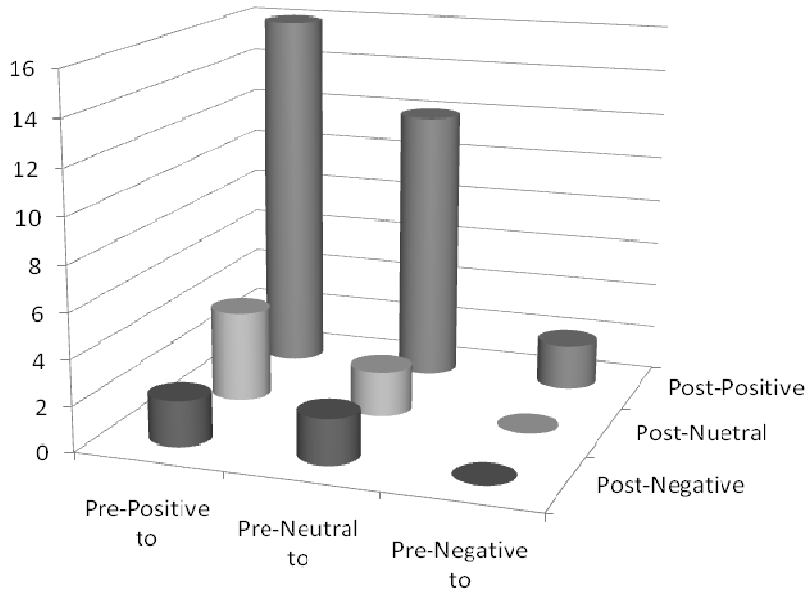

Figure 2. Change trends from pre to post-moderation.

The tables above show the spread of coded data using the descriptor categories. Figure 2 represents the numbers of teachers' attitudes that changed or remained constant following their experiences of moderation. These figures show a trend towards a positive attitude however Figure 2 illustrates that a significant proportion changed their attitudes from positive/neutral to neutral/negative or a trend towards the negative attitude. The quantitative nature of these findings provides a simplistic overview of the situation. Although it would appear that teachers are generally favourable of the use of standards and moderation to achieve consistency in teacher judgements from the categorisation statistics of this initial analysis process a more detailed qualitative analysis of interview data offers further insights.

Teachers' positive views

Interviewees agreed that consensus was achieved through the use of standards and moderation. Broadly speaking, the use of standards to assess students' work was seen as effective in supporting teachers' efforts to be fair markers. Comparing individual teachers' assessments with other teachers' assessments was frequently reported as a powerful and efficient process to reconcile results and thus was considered a fair and appropriate process. These responses represent a positive attitude towards the use of standards and moderation and

the underlying theme of consensus, or consistency of teacher judgement, becomes apparent from an analysis of the transcripts. For example:

> "And it's nice to be able to talk to your peers and see that we've all, we all, sort of, come to the same spot…"

> "People agreed that our decisions were, were the same, very similar to the decisions they'd made. And, um, I guess coming to the district moderation, um, able to see that as a cluster, as well, our decisions were very similar and, to a lot of the other schools in the district so, yep, so that was very positive."

Another identified theme related to peer support. Teachers viewed collaborating with peers as beneficial and constructive for peer relationships, individual confidence building and developing notions of belonging and acceptance. These important aspects of teacher well-being and peer support are qualities that educational leaders encourage. This is a highly beneficial aspect of the moderation process and a clear indication of teachers' approval and support for this process.

> "…so you come out of it as a stronger, as a stronger team."

> "Um, I think, I think they [moderation meetings] are necessary, as I have mentioned before, practically it is very difficult to moderate at regular intervals, but I certainly do see the benefits in moderation. Um, especially with newer teachers. But I mean, even more experienced teachers, I think collaboration is vitally important."

Fairness and equity were themes identified from the interview data analysis that related to the assessment of student achievement. Teachers' saw the use of standards and moderation processes as fair and equitable assessment processes. This was considered to be an even-handed approach towards the assessment of all students' work and teachers opined that a fair and even arena for assessment was possible with the elimination of biases.

> "Without referring back to the standard you find yourself way off course and it's not fair to the kids in the end."

> "…and then, of course, there is the student side and the consistency of marking will assist these students in whether they travel to other schools and also there is a fairness and a consistency of judgement with other schools, as well."

Teachers considered the use of standards and moderation were sound practice with the expressed assumption that teacher consistency of judgement was a facet of this process. From different interviews a range of comments illustrating this view were analysed.

> "I just think it's the whole sharing with each other and that's professional, it's good professional development."

> "It was a really good process. I think the conversation bit of the process was the most valuable part…"

> "I just think it's good to have someone else to look at it and give me their feedback."

The major themes derived from an analysis of data categorised as positive attitude included notions of consensus through use of standards and moderation, peer support and equity towards students. Teachers viewed the process of moderation using standards as a sound and thorough process with the underlying understanding that this process does achieve consistency of teacher judgement. This is indicative that the majority of teachers believe that moderation using standards does result in consistency of teacher judgement.

Teachers' neutral views

As indicated in the earlier figures the majority of neutral attitude descriptors were coded at the pre-moderation stage. This is because a large proportion of these teachers had no previous experience of using standards in moderation. Several indicated a neutral response by articulating very little in regard to the concept and consequently there were very few qualitative descriptions of teachers' neutral attitude. The excerpts that follow are typical of those coded as neutral.

> (Interviewer) "What are some of the factors or influences that you consider need to be taken into consideration to ensure consistency in teacher judgement?"
> (Teacher 1) "I'm not really sure..."
> (Interviewer) "Have you had much experience with moderation?"
> (Teacher 1) "No."
> (Interviewer) "No? Okay. So what are you hoping to gain from your participation in moderation this afternoon?"
> (Teacher 1) "Um, learn a bit more about it. See how other people think, um, why they might have chosen to give the marks they did and, um, yeah, just gain more experience, really."

> (Interviewer) "Have you had experience with moderation before?"
> (Teacher 2) "Not for, not personally. Only second-hand experience."

> (Interviewer) "Now have you had experience with moderation before?"
> (Teacher 3) "Not really, no. No, no. Do you mean with other schools?"

Some teachers felt that the moderation process was inadequately designed and did not function at an acceptable level. They indicated they had problems with the conceptual quality of the QSA materials in terms of lack of detail of the written standards or perceived ambiguity in the standards descriptors that impacted on the accuracy of assessment. The problems related here are not concerned with notions of standards in moderation but that the materials provided were not adequate. To illustrate:

> "There were some answers that really didn't fit into the actual criteria, um, and that caused a little confusion in where you put them. Um, there were some criteria that really didn't match the question at all, um, but it gave us a much better idea of making sure that everything was equitable. However, having had a brief look at some of the other schools, um, they have not interpreted it the same way as we have, so perhaps there needs to be a little bit clearer definition."

Some teachers found responding to questions about standards and moderation processes difficult. Possibly, they had not fully grasped the concept of moderation or were confused.

The response below gives an indication of the imprecise nature of some interviewee responses that were neither positive nor negative and so have been coded as neutral.

> (Interviewer) "Do you consider that teachers would benefit from guidelines for applying standards to the assessment evidence that they collect?"
> (Teacher 4) "Um, I suppose it depends on what you mean by guidelines. If it's too specific it might be too hard but then again if it's too vague it might be too hard as well, so, sort of, I guess it just gets, it's whatever works well for the teacher, I suppose. Somewhere in the middle would probably be the best, but that's a bit vague as well, so..."

Teachers participated in the moderation exercises without adequate preparation for the tasks they were asked to perform. This circumstance was attributable to a number of factors including the amount of curriculum and assessment reform that was being introduced simultaneously by QSA. The lack of preparation and the numerous challenges facing teachers were factors that prevented them from making decisions about the efficacy of the moderation processes. Consequently excerpts such as those that follow were coded as neutral.

> "I suppose … looking back with hindsight, we would have preferred to sit down and actually have a really good go over the standards beforehand and we hadn't had time to do that."

> "Ah, probably I would have needed to – and I will next time – is read that level B and C closer, you know, because the words were 'or' or 'and' in the end, and it took a few scripts to get to realise what the difference is."

The analysis of the neutral coded responses reflects an absence of either a positive or negative attitude for some teachers. A range of thinking is represented as neutral and it is understood that the absence of a clearly stated opinion does not necessarily mean that the interviewees' opinions are categorically neutral. However, it was decided that the neutral descriptor as outlined earlier would include those opinions that could not be readily categorised. In this way these voices were incorporated and represented in the most logical categorisation of these views.

Teachers' negative views

The negative attitude descriptors were identified in only nine interview transcripts. The views categorised as negative were lengthy, committed opinions and well articulated concerns. The key themes are now presented and although the lengthy conversations with the interviewer cannot be presented in full some discussions are contextualised for meaning.

It was suggested that using standards and moderation was unfair as students from some geographical areas were disadvantaged because of the social differences in these contexts. These teachers indicated that the prevailing societal issues hampered the possibility of all students receiving an equitable education and any attempt at moderation across the social strata could not deliver a fair and equitable outcome. This is an important point that has implications for the validity of the assessment practice. Teachers in certain regions need to draw on particular pedagogical and curriculum understanding and skills to attend to the diverse range of student needs in these locations. That is, there is a need for "pedagogical-assessment fit" at local and system levels and that such fit supports student learning and

successful completion of assessment requirements. A degree of flexibility is also required in the curriculum to address the regional and cultural differences that exist. Without adequate training and resources, standards and moderation practice alone will be insufficient.

> (Interviewer) "So, what are some of the factors or influences that need to be considered to ensure consistency of teacher judgement?"
> (Teacher 5) "That's a really, really messy question. And it's really messy because, ah, the background knowledge of kids, regardless of their science teaching over a continuum of years, varies greatly from kids in Cunnamulla to kids in Ascot, and so to turn around and say that, ah, the teachers in these environments can mark consistently is very, very difficult as well. Now, I really don't think that moderation is going to cure that, that issue and, ah, to turn around and say this kid in Cunnamulla knows more or less than this kid in Ascot is, I don't really think, what the issue's all about. …Um, and that, the background knowledge of these kids isn't necessarily indicative of what they've been taught or how they've been taught and so how do we ensure that the teachers are marking the same way? I really don't think you can."

Another theme identified as negative illustrates that, for these teachers, the process was too complex, too time consuming, requiring too much effort from teachers and administrators, and too disruptive to school routines. These teachers did not believe that the effort justified the ends and any gains of the process were deemed insignificant. This is another important point that reflects the need for adequate training and resourcing in the implementation of major curriculum and assessment reforms such as experienced by these teachers.

> (Interviewer) "Anything else? Any other comments you wanted to make?"
> (Teacher 6) "Um, I think we've got to really look at why we're doing this, and again it might sound harsh, but if the bottom line is to make sure that, um, a B for a Weipa student is the same for a Brisbane student, is that overly relevant to a 14 year old year nine student? If it is, fine, do it. But look at the time that it takes."

One teacher expressed concern about the consistency of interpretation of the standards. Teachers apparently did not uniformly agree on the intended meanings of the standard descriptors and therefore it was suggested that they would not be consistent in their judgements.

> "…the biggest problem we had was actually the, our interpretation of the standard and how it applied to the question, …so some teachers would interpret it that the student had understood the fundamental concepts, whereas the descriptors didn't allow for that. Descriptors said we are looking for these specific concepts and we could go through where they say, 'Alright, I think this person has demonstrated to me that they understand the concept.' We go through the descriptor and think, 'Well, they didn't explain it this way so that's actually a C or a D,' even though the classroom teacher thought, 'I think she explained it well enough to get an A or a B.'"

Teachers seemed a little uncertain about the process of using standards in moderation. For example in one instance a teacher described how he achieves consistency across his own range of students' marks and not how a group of teachers can achieve consistency across schools, students, subjects and marks using the standards in moderation. This seems to be a misunderstanding of the intention of the moderation process for this particular teacher

favouring his own method, discounting the value of discussion and negotiation with the group.

Another teacher suggested that the teachers in the moderation could agree on the interpretation of the standards without considering them in great detail. She maintained that a more in depth knowledge of the standards would be beneficial to the marking process and to consistency of teacher judgment. This teacher thought the standard descriptors were inadequate and did not reflect grades realistically. Teachers used grades such as D- or C+ to represent the qualities of student work not represented in the standard as formulated in the QSA materials.

A further critical comment from another teacher was that the moderation system and the use of standards did not allow for the assessment of students' effort. Reference was made to a student who had written extensively and well, yet had failed to address the necessary point of the question. This teacher felt that failing this student was unfair and discriminatory.

The themes associated with the negative descriptor are varied and include concerns about the use of standards in moderation and doubts about how these processes can achieve consistency in teachers' judgement. The four main themes are: first, that using standards does not allow for the social differences amongst students' cultural contexts inherent in Australian society. Second, differences in students, teachers, teaching approaches and curriculum choices will ultimately produce different responses to common assessment tasks, thus the grades awarded using standards is an inequitable process. The third theme relates to the quality of teachers work-lives and disruptive factors to school routines. That is the process of using standards in moderation is too much work for too little reward. Fourth, consistency of interpretations of the QSA materials (*Teacher Guidelines* and the *Guide to Making Judgments*) varied among teachers and it was suggested that individual readings and this lack of consensus would impact on the consistency of teachers' judgements.

ICT-moderation process

This study found that there was no qualitative difference in participants' views about the use of standards in moderation and consistency of teachers' judgements between those who participated in ICT-moderation and those who did not. Teachers' views were qualified by their attitudes not by the mode of moderation. Although the teachers who took part in the ICT-moderated exercises can be separated out by unique codes, as seen in Table 4 and Figure 2, their interview transcripts were analysed together with the others of this data set. Of the 7 teachers who participated in ICT-pre and ICT-post-moderation interviews none mentioned the ICT process as a contributing factor to their attitudes and beliefs concerning the use of standards in moderation as a means of producing consistency in teachers' judgements.

The teachers had a positive response to the ICT-moderation mode with most experiencing little or no difficulty using the system. A major advantage of the process was seen to be the ability to overcome *the tyranny of distance* so frequently reported by teachers in remote Australian schools. A process that provided teachers with the means to join an online group of peers and carry out moderation was seen as a valuable and productive exercise.

> "…so I think, yeah, it's certainly very valuable for – especially for those schools where they're not, or they can't get that moderation process at the school level – but

> even when you do, I think it's really good, because we just don't talk to other schools in other regions, it just doesn't happen, so I think it's really good to have that."

Another point that was observed by at least two teachers was the provision for other teachers, from their own school or department, to take part as observers or panel members, by simply being present at the time of the ICT-moderation exercise. In this way several teachers could attend the one computer terminal and observe or participate in moderation. This was seen as a potentially valuable tool for training and alignment of consensus between teachers within schools.

The main advantage of the ICT-moderation process was the opportunity for those teachers who would normally never participate in moderation to do so with teachers from other schools. The realisation that the process can take place in remote areas, or even in metropolitan areas, with a relatively small degree of organisation, cost and time expenditure was seen to be very beneficial for teachers and HODs.

Nevertheless the main concerns related to the efficiency of the ICT process and perceptions about preparation for events, use of technology, aligning schools' schedules, time taken for scanning or photocopying requirements, and the efficiency and reliability of the network system. One other concern was the notion that online moderation does not allow for a sense of community and intimacy in discussions as available in the face-to-face mode. The opportunity to see body language and experience the meeting were cited as integral to developing a community that could produce greater meaning and consistency between participants as possible in the face-to-face mode. Although these were not advantages of the online mode this concern was limited with the benefits of ICT moderation being conveyed in stronger terms.

Conclusion

This research set out to study teachers' perspectives of the use of standards and moderation as a means to achieving consistency of teacher judgement. While many participating middle years teachers believed that the use of standards in moderation produces consistency in teachers' judgements, important differences exist among the teachers in this study. Those in favour of the use of moderation in national and state assessment reforms reported that such a system should produce consistency of teacher judgement. These teachers valued the opportunity to interact with others to determine the comparability of their judgements. Participation in moderation was viewed to be reassuring as familiarity with the standards and confidence with the procedures increased. A large number of participants who were not committed to moderation and the use of standards, or were insufficiently informed about them , preferred to remain neutral in their attitude. Many of these teachers had very little or no prior experience of moderation previously. Their neutral views were attributable to a lack of preparation for the exercise, lack of training and familiarity with the materials, processes and procedures and limited understanding of the purposes of standards-referenced moderation.

Variability in teachers' attitudes towards the use of moderation to achieve consistency relates to their particular context. This includestheir familiarity with the standards, moderation practices, curriculum, assessment tasks, students and location of the school. A range of factors is also influential in how judgements are reached – the teachers' notions of quality, their professional experience and understanding, syllabus use, their own assessment beliefs, attitudes and practices. As identified by others (Davison, 2004; Cooksey et al., 2007; Sadler, 2009) context remains an important determining factor in the validity and reliability of

assessment and moderation practices. Teachers will draw on their tacit knowledges (e.g., teachers' personal knowledge of students; knowledge of curriculum and teaching contexts; prior evaluative experience and individual tacit knowledge of standards not elsewhere specified) for judgement purposes. In this analysis of teachers' perspectives it became apparent that other knowledge can be used as a reason for discounting, or even subverting the stated standards and the use of moderation practices (Wyatt-Smith, Klenowski and Gunn, 2010).

Those teachers who expressed negative attitudes towards the idea of moderation and the use of standards to achieve consistency in teachers' judgements raised important issues again related to context, and in particular, cultural and regional differences. To address the assessment purposes of validity and fairness it is important that policy officers ensure that the resources, materials and procedures make provision for such differences. The lack of alignment between the introduction of a new curriculum (ELs) and assessment format and procedures (standards and moderation) also needs to be addressed when teachers are confronted with major reforms simultaneously. This study highlights how teachers were involved in a trial, with materials that were being developed which required a steep learning curve in some cases with limited professional development.

Teachers' involved in ICT-moderation did not express different views from those who participated in face-to-face moderation. However, despite the small sample size the reported benefits of using ICT-moderation appear greater than the reported concerns with this mode of moderation.

The key implications for policy and practice that derive from this study are that it is crucial that guidelines and professional development opportunities be provided to teachers about desired judgement practice and the legitimacy (or otherwise) of the various resources available for teachers to draw upon. The use of standards and moderation practice presents new content, a different assessment format, different processes and procedures that challenge the confidence and their current status as experts. Confronted with competing demands it is vital that the particular context of teachers is considered in efforts to promote teachers' professional learning, develop their assessment capacity and provide them with an opportunity to participate in system level standards-referenced reporting.

Although some teachers were sceptical of moderation practices involving defined standards, or qualified their support, it is clear that teachers generally believe that this type of system, used at national or state levels is an appropriate assessment initiative to address issues of equity and comparability of teacher judgement. However, teachers also recognise and suggest that the moderation system can be improved through further research and development efforts. Such efforts are motivated by recognition that quality education for all students requires teachers to have a repertoire of assessment practices that are valid and reliable and that take seriously the challenge of "pedagogic-assessment fit".

Acknowledgements

The authors wish to acknowledge that the project from which these ideas are derived was funded by the Australian Research Council Linkage Program and we wish to further acknowledge the involvement and support provided by our Partners, the Queensland Studies Authority (QSA), the National Council for Curriculum and Assessment of the Republic of Ireland and Queen's University, Belfast, Northern Ireland.

We wish to also acknowledge the very important contributions made by other researchers on the project including Lenore Adie (Queensland University of Technology), and Stephanie Gunn, Peta Colbert and Rachelle Wyatt-Smith (Griffith University).

References

Adie, L. (2010) Developing Shared Understandings of Standards-based Assessment: Online Moderation Practices across Geographically Diverse Contexts, unpublished PhD thesis, QUT, Kelvin Grove.

Bazeley, P. (2007) *Qualitative data analysis with NVivo* (Los Angeles, SAGE).

Creswell, J. (2008) *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (3rd ed.) (Upper Saddle River, N.J, Pearson/Merrill Prentice Hall).

Cooksey, R., Freebody, P., & Wyatt-Smith, C. (2007) Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing, *Educational Research and Evaluation,* 13(5), 401-434.

Davison, C. (2004) The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools, *Language Testing,* 21(3), 305-334.

Klenowski, V. (2006) *Evaluation report of the pilot of the 2005 Queensland Assessment Task (QAT).* Available online at: http://www.qsa.qld.edu.au/research/reports.html (accessed 17 January 2010).

Klenowski, V. (2007) *Evaluation of the effectiveness of the consensus-based standards validation process.* Available online at: http://education.qld.gov.au/corporate/newbasics/html/lce_eval.html (accessed 2 February 2010).

Klenowski, V., & Adie, L. (2009). Moderation as judgment practice: Reconciling system level accountability and local level practice, Curriculum Perspectives, 29(1), 10-28.

Lincoln, Y. S., & Guba, E. G. (1985) *Naturalistic Inquiry* (Beverly Hills, SAGE).

Maxwell, G. (2007) *Implications for moderation of proposed changes to senior secondary school syllabuses.* Paper commissioned by the Queensland Studies Authority (Brisbane, Queensland Studies Authority).

Maxwell, G. S. (2009) Defining Standards for the 21st Century, in C.M. Wyatt-Smith & J. J. Cumming (eds.), *Educational Assessment in the 21st Century, (*Dordrecht, The Netherlands, Springer)

Miles, M., & Huberman, A. M. (1994) *Qualitative data analysis: an expanded sourcebook* (Thousand Oaks, Sage).

Queensland Studies Authority. (2005) *Moderation processes for senior certification*. (Brisbane, Australia, Queensland Studies Authority).

Queensland Studies Authority. (2007) *Information statement February 2007: Essential learnings draft 2*. (Brisbane, Australia, Queensland Studies Authority).

Sadler, D.R. (2009) Indeterminacy in the use of preset criteria for assessment and grading in higher education, *Assessment and Evaluation in Higher Education*, 34(2), 159-179.

Sadler, D. R. (2005) Interpretations of criteria-based assessment and grading in higher education, *Assessment and Evaluation in Higher Education*, 30(2), 175-194.

Saldaña, J. (2009) *The coding manual for qualitative researchers* (London, SAGE).

Simons, H. (2009) *Case study research in practice* (Los Angeles, SAGE).

Wyatt-Smith, C. M. (1999) Reading for assessment: How teachers ascribe meaning and value to student writing, *Assessment in Education*, 6(2), 195-223.

Wyatt-Smith, C., & Castleton, G. (2005) Examining how teachers judge student writing: an Australian case study, *Journal of Curriculum Studies*, 37(2), 131-154.

Wyatt-Smith, C., & Castleton, G. (2004) Factors affecting writing achievement: Mapping teacher beliefs, *English in Education*, 38(1), 37-61.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. J. (2010) The centrality of teachers' judgment practice in assessment: a study of standards in moderation, *Assessment in Education*, 17(1), 59-76.

[i] Readers interested in details of the project are advised to go to http://www.griffith.edu.au/education/standards-driven-reform).